CONGESTION CONTROLLER FOR ETHERNET SWITCH

BACKGROUND OF THE INVENTION

5       1.    Field of the Invention

The present invention relates to a congestion controller, and in particular, it relates to a congestion control system to control transmission traffic using a PAUSE frame when congestion occurs in an environment in

10    which a plurality of Ethernet switches are connected.

      2.    Description of the Related Art

Conventionally, if congestion occurs in an Ethernet (Registered Trademark) network, a PAUSE frame defined by IEEE 802.3 is sent from a switch such as a

15    switching hub, in which the congestion occurs, to inner switches, to thereby control input traffic (See Note 1).

Fig. 1 shows an example of congestion control using a PAUSE frame, and Fig. 2 shows a PAUSE frame format.

20         In Fig. 1, if congestion occurs in a link "α" on the transmission port side of a switch 1 (SW1), the congestion is detected by a threshold value of a buffer (queue) in the switch 1. Thereafter, in order to restrict the transmission traffic from inner switches 2-5

25    (SW2-SW5), the switch 1 sends a PAUSE frame to the switches 2-5. As shown in Fig. 2, the PAUSE frame sets, as the destination address (DA: Destination Address), a multicast address representing the PAUSE frame.

In a parameter field of the PAUSE frame, a

30    timer value of Pause time, e.g., XX (ms) is set. If the timer value is 0, it represents that the PAUSE state is released and the transmission state is established. Other fields are the same as those of a typical Ethernet frame and are not discussed herein.

35        The inner switches 2-5 suppress traffic transmission to the outer switch 1 with which the congestion has occurred, for a predetermined time, based

on the timer value included in the received PAUSE frame.
If the node which receives the PAUSE frame is a terminal,
transmission is suppressed temporarily based on the timer
value, as long as the terminal has a NIC (Network
Interface Card) which supports the PAUSE frame.  When the
PAUSE time is up, the transmission which has been
suspended is resumed.

Note 1: Okabe Yasuichi <Detailed TCP/IP
Protocol 9th Ethernet (No. 4) Flow Control and VLAN,
Troubleshooting 1. Flow Control of Ethernet> [online]
October 2nd, 2001, Network Technology Lecture, [Searched
on September 9, 2002], Internet <URL:
http://www.atmarkit.co.jp/fwin2k/network/tcpip009/tcpip01
.html>

As described above, the PAUSE frame defined by
IEEE 802.3 causes the switch 1 with which the congestion
has been detected to request the inner switches 2-5 and
lower terminals thereof to suppress transmission of all
the traffic to the switch 1 for a predetermined time.
However, such conventional operations have problems
discussed below.

(1)  Reservation of QoS (Quality of Service)

Because a PAUSE frame does not distinguish
the kinds of traffic, e.g., audio and/or video traffic
which have strict requirements on delay time and jitter,
etc., and data traffic such as FTP (File Transfer
Protocol) which has less strict requirements are treated
equally.  In particular, the QoS for the former traffic
cannot be reserved.

(2)  Improvement of Effective Throughput

Because transmission of all the traffic to
the switch 1 from the switches 2-5 and lower terminals
thereof is simultaneously suppressed for a predetermined
time, there is a possibility that the PAUSE times lapse
concurrently, so that congestion state and non-congestion
state repeatedly appear.  To improve the throughput, it
is preferable that the amount of traffic be constant

without fluctuating.  The effective throughput cannot be improved by the conventional congestion control.

(3)  Equality of Traffic

All the flows are stopped without identifying which flow causes the congestion, it is possible to allocate bandwidth equally to a plurality of existing flows. Namely, transmissions of traffic other than the traffic with which the congestion has occurred are equally suppressed.

SUMMARY OF THE INVENTION

The present invention is, in view of the above problems, aimed to provide a congestion controller in which

when the transmission traffic is controlled in accordance with a PAUSE frame in the event that congestion occurs in an environment in which a plurality of Ethernet switches are connected, the attributes of the traffic are analyzed, whereby the improvement of QoS and the effective throughput, and equality of each traffic can be achieved.

According to the present invention, there is provided a congestion controller for an Ethernet switch comprising

a plurality of transmission queues which have different priorities,

a receiving means for receiving a PAUSE frame,

a suppression means for suppressing transmission traffic from the transmission queues by the received PAUSE frame, wherein

the suppression means suppresses the transmission traffic from a transmission queue of the lowest priority by the PAUSE frame received at a time other than the PAUSE time, and restricts the transmission traffic from the transmission queue of the highest priority, by the PAUSE frame received during the PAUSE time.

According to the present invention, there is

provided a congestion controller for an Ethernet switch comprising

a transmission queue,

a receiving means for receiving a PAUSE frame,

5    a shaping means for shaping the transmission traffic from the transmission queue by the received PAUSE frame, wherein

the shaping means restricts transmission speed of the transmission traffic from the transmission queue

10   to or below a transmission speed based on a predetermined shaping value.

According to the present invention, there provided a congestion controller for an Ethernet switch comprising

a transmission queue,

15   an identifying means for identifying an input port which causes congestion by counting packets resident in the transmission queue corresponding to the input port,

a transmission means for transmitting a PAUSE

20   frame to other switch which is connected to the identified input port.

BRIEF DESCRIPTION OF THE DRAWINGS

The present invention will be more clearly understood from the description as set forth below with

25   reference to the accompanying drawings.

Fig. 1 shows an example of a conventional congestion control by a PAUSE frame.

Fig. 2 shows an example of a format of a PAUSE frame.

30   Fig. 3 shows an example of a network construction to which the present invention is applied.

Fig. 4 shows a first embodiment of the present invention.

Fig. 5 shows an example of a control flow in Fig. 4.

35   Fig. 6 shows a second embodiment of the present invention.

Fig. 7 shows an example of a control flow in Fig. 6.

Fig. 8 shows an example (1) of a shaping operation in Fig. 6.

Fig. 9 shows an example (2) of a shaping operation in Fig. 6.

Fig. 10 shows a third embodiment of the present invention.

Fig. 11 shows an example (1) of a control flow in Fig. 10.

Fig. 12 shows an example (2) of a control flow in Fig. 10.

Fig. 13 schematically shows an example of a traffic analyzing process.

DESCRIPTION OF THE PREFERRED EMBODIMENTS

Fig. 3 shows an example of a network construction according to the present invention.

In this embodiment, the structure is basically the same as that shown in Fig. 1 and described above, each of, for example, switches 2 and 6 which are connected to one of the receiving ports of the switch 1 and each of switches 3, 7, 8 etc., which are connected to other receiving ports configure a transmission unit of a PAUSE frame multicast.  In this example, congestion occurs at a link "α" of a sending port of the switch 1.

Fig. 4 shows a first embodiment of the present invention and Fig. 5 shows an example of a control flow shown in Fig. 4.

In this embodiment, a congestion control is performed to reserve QoS.  First, the switch 1 of which the congestion is detected based on a threshold value or the like of a buffer sends a PAUSE frame to the inner switches 2-5 (S101 and S102).

The PAUSE frame is detected by a CPU 21 in each of the switches 2-5 (S103), and is transmitted to a scheduler 22.  The scheduler 22 controls transmission queues 23-25 having high, middle, or low priority to thereby control the transmission traffic based on the priority.  For example, the high priority is assigned to

video and/or audio data which requires real-time
processing, and the low priority is assigned to data
transmission such as FTP.

The scheduler 22 determines the transferred PAUSE
frame. If the PAUSE frame is received at a time other
than the PAUSE time (normal transmission period),
transmission from the low priority queue 25 is controlled
(S104 and S105). Another PAUSE frame is received during
the PAUSE time, transmission from the middle priority
queue 24 is controlled (S104 and S106).

When the timer of the PAUSE time is up, or a PAUSE
completion notification (the timer value of the PAUSE
frame = "0") is received, transmission of the suspended
queue begins (S107 and S108). As described above,
according to the present invention, the PAUSE operation
in consideration of the priority can be carried out. As
a result, video and/or audio traffic of the high priority
are not be disrupted and the QoS thereof can be reserved,
even if the PAUSE frame is received.

Fig. 6 shows a second embodiment of the present
invention, and Fig. 7 shows an example of a control flow
of Fig. 6.

In this embodiment, a congestion control in which
the effective throughput is improved by shaping is
performed. First, the switch 1 of which congestion has
been based on by a threshold value or the like of a
buffer sends a PAUSE frame to inner switches 2-5 (S201
and S202).

The PAUSE frame is detected by the CPU 21 in each of
the switches 2-5 (S203), and is transferred to a shaper
31. The shaper 31 starts the shaping operation to
restrict the transmission speed of transmission queue 32
to 50% of the original physical speed (S204).

Fig. 8 and Fig. 9 show examples of the shaping
operation shown in Fig. 6.

In an example of Fig. 8, a gap is calculated based
on the frame length. Namely, a gap based on the frame

length of each transmission frame and the shaping rate
value (50%) is calculated and obtained by the following
formula:

Gap value[sec]=frame length[sec]×(100-shaping

5    rate[%])÷100

As a result, a gap of an identical frame length is
added to each transmission frame from the transmission
queue 32, and thus, the shaping operation in which, for
example, the transmission speed is limited to 50% of the

10   physical speed is performed. In an example of Fig. 9,
the time $\Delta t$ for which the frame transmission normally

lasts is reduced to $\Delta t$ /2, so that the shaping is

achieved by 50%.

After that, when the timer of the PAUSE time expires

15   or the PAUSE completion notification (the timer value of
the PAUSE frame = "0") is received, the shaping operation
ends and normal data transmission is resumed (S205 and
S206). In this example, the shaping values (%) of the
inner switches 2-5 are pre-determined to prevent

20   congestion at the switch 1.

Namely, the shaping value (%) of each inner switch
2-5 is determined so that a sum of the effective
transmission speeds of the inner switches 2-5 does not
exceed the effective transmission speed of the switch 1.

25   As described above, in this example, the switches 2-5 can
continue the transmission even if the PAUSE frame is
received, and accordingly, a congestion control in which
the resultant effective throughput of the whole network
is achieved can be achieved.

30   Alternatively, this example can be combined with the
first embodiment, wherein for example, the shaping
operation is not performed for the high priority queue
23, or the shaping degree is increased to 80% instead of
suppression of the transmission of the low priority queue

35   25 to thereby reserve the QoS and enhance the effective
throughput improvement.

Fig. 10 shows a third embodiment of the present invention. Fig. 11 and Fig. 12 show an example of a control flow of Fig. 10.

In this embodiment, a congestion control is
5 performed to achieve equality of the traffic by allocating the bandwidth equally to a plurality of the existing flows. In contrast to the first and second embodiments applied to the inner switches 2-5, this embodiment is applied to the switch 1.

10 In Fig. 10, the switch 1 detects the congestion when the transmission queue 42 of the sending port to the link "α" exceeds a predetermined threshold value (S302). Thereafter, in this embodiment, attribute of the packet (e.g., sending address or port number of TCP/UDP)
15 resident in the transmission queue 42 is analyzed (S302). The detailed flow of this congestion factor traffic analysis routine is shown in Fig. 12.

In Fig. 12, first, the senders' addresses of the X packets (for example, X=100) in the late input frame are
20 checked, and corresponding input ports are retrieved with reference to a table 43 (Fig. 10) comprised of a learning table or a routing table of the MAC address, based on the sender's addresses (S401 and S402). Next, the number of the received packets is counted for each searched input
25 port, by a counter 41 (Fig. 10). Thus, the input port in which the count value is largest can be specified (S403 and S404).

Fig. 13 schematically shows an example of the above analyzing process.

30 (a) in the drawing shows an operation when the number of the frames (packets) resident in the transmission queue 42 is below the threshold value. When the number of the frames reaches the threshold value as shown in (b) in the drawing, in this example, the
35 senders' addresses of the preceding 100 frames including the frame by which the number of the frames becomes the threshold value are checked, and the corresponding input

ports are retrieved with reference to the table 43.

It is obvious from Fig. 10 that marks ○, □, etc., represent frames from different input ports. Accordingly, the respective numbers of the frames ○, frames □, etc., are counted by the respective counters 41. For example, if the number of the frames ○: frames □: frames ..., are l, m and n, respectively (wherein l>m, n, it is judged that the input port of the frames ○ is the cause of congestion, and is subject to the congestion control (S403 and S404).

In the above example, an input port of the largest number of frames is determined to be cause of congestion, but it is also possible to judge that an input port of the specific frames corresponding to those not less than 50% of the total resident frames is the cause of the congestion. Alternatively, a predetermined number of frames is continuously monitored, so that the flow which causes the congestion can be determined from the information of the resident frames at the time when the congestion is detected.

Referring to Figs. 10 and 11, the switch 1 sends a PAUSE frame only to an input port of the traffic (frame ○) which has been identified as the cause of congestion (S303). In the example of Fig. 10, the PAUSE frame is sent to the switch 2 connected to the above input port. Accordingly, only the switch 2 which has received the PAUSE frame starts the PAUSE process and the control ends (S304-S308).

Moreover, the attributes of the packet (e.g., sending address, port number of TCP/UDP, etc.) whose traffic has been specified as the cause of congestion are preset in the PAUSE frame, so that the switch 2 which receives the PAUSE frame can suppress transmission of only that traffic. In this alternative, the switch 2 must be provided with function to discriminate the specific attribute and dynamically restrict the corresponding traffic.

As described above, in this example, as only a switch specified as the cause of congestion becomes an object for the PAUSE process, it is possible to prevent only the switch from occupying network resource, resulting in equal allocation of the bandwidth to other existing switches.

This embodiment is applied to the switch 1 which detects the congestion and can be combined with the first and the second embodiments applied to the inner switches 2-5 to carry out a congestion control in which the QoS is reserved, the effective throughput is enhanced, and equality of the traffic is ensured.

As described above, according to the present invention, a congestion control considering the QoS can be performed. Moreover, according to the present invention, no repetitive appearance of the congestion state and non-congestion state occurs, and the effective throughput in the network in which the congestion has occurred can be improved. Furthermore, according to the present invention, only a specified transmission traffic which causes the congestion is restricted, whereby equal utilization of the network resource can be ensured.